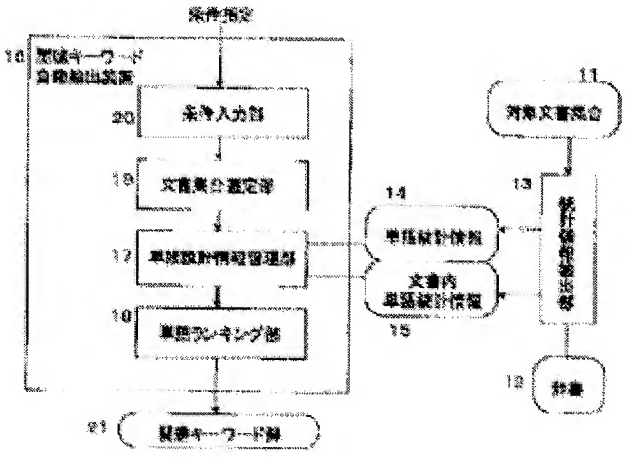
AUTOMATIC EXTRACTION DEVICE FOR RELATIVE KEYWORD, DOCUMENT RETRIEVING DEVICE AND DOCUMENT RETRIEVING SYSTEM USING THESE DEVICES

Publication number: JP11025108
Publication date: 1999-01-29
Inventor: SATO MITSUHIRO; NOGUCHI NAOHIKO; SUGANO YUJI; NOMOTO MASAKO; INABA MITSUAKI; FUKUSHIGE TAKAO
Applicant: MATSUSHITA ELECTRIC IND CO LTD
Classification:
- international: ***G06F17/30;*** **G06F17/30;** (IPC1-7): G06F17/30
- European: G06F17/30T1E
Application number: JP19970176822 19970702
Priority number(s): JP19970176822 19970702

Also published as:
EP0889419 (A2)
US6212517 (B1)
EP0889419 (A3)
CN1206158 (A)
EP0889419 (B1)
more >>

Abstract of **JP11025108**

PROBLEM TO BE SOLVED: To automatically extract a relative keyword which is matched with the characteristics of a document to be practically retrieved and which is capable of obtaining one or more retrieval results at the time of executing retrieval using the keyword. SOLUTION: An automatic extraction device for relative keywords is provided with a document set selection part 19 for specifying a partial set of each document based on the attribute information, input retrieval expression, etc., of the document, a word statistic information management part 17 for managing the statistic information of respective words in the whole objective document 11 and words appearing in each document as well as their statistic information 15; and a word ranking part 18 for calculating the importance of each word appearing in a partial set of a certain document and for aligning respective words in the order of importance, wherein the management part 17 quickly finds out the statistic information of respective words in the whole document and a specified partial set of the document. Consequently, words appearing in a certain document set can be ranked based on their importance and a part of the ranked words can be presented as a relative keyword.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号

特開平11－25108

(43)公開日　平成11年(1999)１月29日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３１０Ａ
		15/40	３７０Ａ

審査請求　未請求　請求項の数17　ＯＬ　（全 17 頁）

(21)出願番号　特願平9－176822

(22)出願日　平成９年(1997)７月２日

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地

(72)発明者　佐　藤　光　弘
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　野　口　直　彦
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　菅　野　祐　司
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(74)代理人　弁理士　蔵合　正博

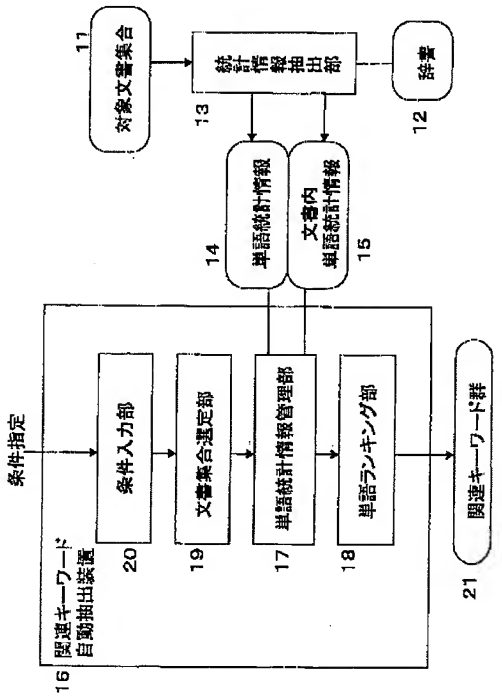
最終頁に続く

(54)【発明の名称】　関連キーワード自動抽出装置、文書検索装置及びこれらを用いた文書検索システム

(57)【要約】

【課題】　実際の検索対象文書の特性に即し、かつそのキーワードによる検索を実行した場合少なくとも１件以上の検索結果が得られるような関連キーワードを自動抽出すること。

【解決手段】　関連キーワード自動抽出装置として、各文書の属性情報や入力検索式などに基づいて文書の部分集合を特定する文書集合選定部１９と、各単語の対象文書１１全体における統計情報１４および文書毎に出現する単語とその統計情報１５を管理する単語統計情報管理部１７と、単語統計情報１４、１５を基に、或る文書の部分集合に出現する各単語の重要度を算出して重要度の順に整列する単語ランキング部１８とを設け、単語統計情報管理部により、文書全体、および特定された文書部分集合における各単語の統計情報を高速に求める。これにより、或る文書集合に出現する単語を、その重要度に基づいてランキングし、その一部を関連キーワードとして提示することができる。



1

【特許請求の範囲】

【請求項１】　辞書を用いて対象文書集合の各文書に出現する単語または単語の組の出現頻度や分布などの統計情報があらかじめ抽出されている文書集合に対して、各文書に付与された属性情報やユーザが入力した検索式などに基づいて文書の部分集合を特定する文書集合選定部と、各単語の対象文書全体における統計情報、および各文書ごとの当該文書に出現する単語とその統計情報を管理する単語統計情報管理部と、各単語の全文書および各文書ごとの統計情報を基に、特定された部分集合に出現する各単語の重要度を算出して重要度の順に整列する単語ランキング部とを有し、整列された単語群のうちの特定部分のみについて、単語もしくは単語とその重要度の組を抽出し、これを再利用可能な形で提示するようにしたことを特徴とする関連キーワード自動抽出装置。

【請求項２】　前記構成において、特定された部分集合Ａに対して、これに含まれる部分集合Ｂが文書集合選定部により特定された場合に、部分集合Ａに含まれる文書群に出現する単語の統計情報と、部分集合Ｂに含まれる文書群に出現する単語の統計情報との差分を、部分集合Ｂにおける各単語の重要度に加味することで、部分集合Ｂに出現する各単語の重要度を算出して単語ランキングに反映することを特徴とする請求項１に記載の関連キーワード自動抽出装置。

【請求項３】　文書集合選定部に各文書の重みを付与する機能を設け、特定された文書集合の各文書に含まれる単語の重要度に当該文書の重みを加味することにより当該単語の重要度を算出して単語ランキングに反映することを特徴とする請求項１または２に記載の関連キーワード自動抽出装置。

【請求項４】　対象文書集合全体において出現度合いが高頻度または低頻度である単語をあらかじめ定められた閾値を考慮して除外することにより、再利用の際に有効性の高い単語のみが選別できることを特徴とする請求項１乃至３のいずれかに記載の関連キーワード自動抽出装置。

【請求項５】　単語の長さなどその単語の特徴量に応じて除外のための閾値を変化させることにより再利用の際に有効性の高い単語のみが選別できることを特徴とする請求項４に記載の関連キーワード自動抽出装置。

【請求項６】　単語の出現位置や出現する文脈の情報を管理する出現情報管理部を有し、単語の重要度にその単語の出現情報の種類に応じてあらかじめ定められた重みを加味することにより当該単語の重要度を算出して単語ランキングに反映することを特徴とする請求項１乃至５のいずれかに記載の関連キーワード自動抽出装置。

【請求項７】　単語の品詞など、各単語の属性情報を管理する言語属性管理部を有し、当該単語の属性に応じてあらかじめ定められた重みを加味することにより当該単語の重要度を算出して単語ランキングに反映することを

2

特徴とする請求項１乃至６のいずれかに記載の関連キーワード自動抽出装置。

【請求項８】　抽出された単語同士、またはあらかじめ指定された単語群と抽出された単語との間の文字列としての包含関係を、定められた条件により判定する文字列包含関係判定部を有し、当該単語同士に文字列としての包含関係があると判定された場合に、指定された条件に従って、長単位の文字列のみ、もしくは短単位の文字列のみ、もしくは重要度の高い方の文字列のみ、もしくは短単位の文字列および長単位の文字列と短単位の文字列との差分の双方、のいずれかを選択することにより、再利用の際に有効性の高い単語のみが選別できることを特徴とする請求項１乃至７のいずれかに記載の関連キーワード自動抽出装置。

【請求項９】　単語の品詞など、各単語の属性情報を管理する言語属性管理部を有し、当該単語の属性や、指定された部分集合または文書全体における出現頻度、分布等を考慮することにより、抽出された単語を分類して提示できることを特徴とする請求項１乃至８のいずれかに記載の関連キーワード自動抽出装置。

【請求項１０】　分類された単語群のそれぞれについて、その集合を代表する単語を付与する代表語付与部を設け、分類された単語群を代表する代表語群のみ、もしくは代表語と全ての単語を提示できることを特徴とする請求項９に記載の関連キーワード自動抽出装置。

【請求項１１】　辞書を用いて対象文書集合の各文書に出現する単語または単語の組の出現頻度や分布などの統計情報があらかじめ抽出されている文書集合に対して、文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部と、文書検索部４５において検索された文書について、入力された検索式と文書との間の適合度を計算する文書ランキング部４６とを有して成り、文書ランキング部におけるランキング結果を関連キーワード自動抽出装置へ送付し、また関連キーワード自動抽出装置からフィードバックされた関連キーワードを検索条件入力部へ入力することが可能な文書検索装置。

【請求項１２】　文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部とを有して成り、前記検索条件入力部は、ユーザが検索条件を入力するのに加えて関連キーワード自動抽出装置から送られてきた関連キーワードを検索条件として入力することが可能な文書検索装置。

【請求項１３】　辞書を用いて対象文書集合の各文書に出現する単語または単語の組の出現頻度や分布などの統計情報があらかじめ抽出されている文書集合に対して、文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書

の検索を行なう文書検索部と、文書検索部４５において検索された文書について、入力された検索式と文書との間の適合度を計算する文書ランキング部４６とを有して成る文書検索装置と、
前記文書検索装置に接続された関連キーワード自動抽出装置とから構成され、
前記文書検索装置の文書ランキング部から出力されたランキング結果を関連キーワード自動抽出装置へ送付し、また関連キーワード自動抽出装置から文書検索装置の検索条件入力部へ関連キーワードをフィードバックしてキーワード検索を行なうようにしたことを特徴とする文書検索システム。

【請求項１４】　文書検索装置と関連キーワード自動抽出装置との間には文書集合選定部が設けられ、文書検索装置の文書ランキング部から出力されたランキング結果は文書集合選定部に送付されて文書の特定が行なわれ、前記関連キーワード自動抽出装置４８には、文書集合選定部４７が特定した文書の部分集合が入力されることを特徴とする請求項１３記載の文書検索システム。

【請求項１５】　関連キーワード自動抽出装置には、請求項１乃至１０のいずれかに記載の関連キーワード自動抽出装置が用いられることを特徴とする請求項１３または１４記載の文書検索システム。

【請求項１６】　文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部とを有して成る文書検索装置と、
前記文書検索装置に接続された関連キーワード自動抽出装置とから構成され、
前記文書検索装置の検索条件入力部は、ユーザが検索条件を入力するのに加えて関連キーワード自動抽出装置から送られてきた関連キーワードを検索条件として入力してキーワード検索を行なうようにしたことを特徴とする文書検索システム。

【請求項１７】　関連キーワード自動抽出装置には、請求項１乃至１０のいずれかに記載の関連キーワード自動抽出装置が用いられることを特徴とする請求項１６記載の文書検索システム。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、特定の文書集合から、その文書集合を特徴づける語句をキーワードとして抽出するための関連キーワード自動抽出装置、および前記関連キーワード自動抽出装置を利用した文書検索装置に関する。

【０００２】

【従来の技術】文書検索装置において、ユーザが必要とする文書を得るためには、適切な検索語を利用した検索式を入力する必要があるが、ユーザ自身が適切な検索語を想起し難い、という問題がある。そこで従来、ユーザ

が入力した検索語に対して、関連語辞書などを利用して検索語に関連する語を提示することにより、ユーザの再検索を助ける手法などが取られてきた。しかしながら、こうした手法はあらかじめ静的にさだめられた関連語辞書の性質に依存するため、検索対象となる文書の特性に即した関連語が得られない。また、得られた単語で検索した結果少なくとも１件以上の文書が得られることが保証されない、という欠点があった。

【０００３】

【発明が解決しようとする課題】本発明は前記の課題を解決するもので、特定された文書集合における各単語の出現頻度・分布などの統計情報と、検索対象文書全体における単語の統計情報とを考慮して単語の重要度を算出し、これにもとづいて単語をその重要度によってランキングし、ランクの一部である単語群を抽出することにより、実際の検索対象文書の特性に即し、かつ品質の高い関連キーワード群を高速かつ動的に抽出できる、関連キーワード自動抽出装置を提供することを目的とする。

【０００４】また、前記関連キーワード自動抽出装置から得られた関連キーワード群を利用して検索を実行した場合、少なくとも１件以上の検索結果が得られることを保証する文書検索装置及びこれらを用いた文書検索システムを提供することを目的とするものである。

【０００５】

【課題を解決するための手段】本発明は、上記目的を達成するため、関連キーワード自動抽出装置として、各文書に付与された属性情報やユーザが入力した検索式などに基づいて文書の部分集合を特定する文書集合選定部と、各単語の対象文書全体における統計情報や各文書ごとに出現する単語とその統計情報を管理する単語統計情報管理部と、各単語の全文書または各文書内統計情報を基に、特定された文書の部分集合に出現する各単語の重要度を算出して重要度の順に整列する単語ランキング部とを設け、単語統計情報管理部により、文書全体、および特定された文書部分集合における各単語の統計情報を高速に求めることが可能であり、特定された文書集合に出現する各単語を、その重要度の順に高速にランキングし、その一部を関連キーワードとして提示することができる。

【０００６】さらに、前記構成に加えて、単語の属性情報や文書内の出現位置を管理する手段などを設けることにより、単語の重みを変化させ、あるいはランキング後の単語群から特定の条件を満たす単語を削除することで、抽出される単語群の関連語としての精度を向上させることができ、また、抽出された単語群を、語の属性や統計的性質により分類することで、よりわかりやすい関連キーワード提示を行なうことができる。

【０００７】また本発明は、上記目的を達成するため、関連キーワード自動抽出装置と連携した文書検索装置を含む文書検索システムを構成し、抽出された関連キーワ

ードを入力として再利用することにより、抽出された関連キーワードが対象文書の特性に合ったものであり、かつ検索対象が同一の文書群であるならば、そのキーワードによって検索結果が少なくとも１件以上得られることが保障されるため、効率的かつ容易に再検索を行なうことができる。

【０００８】

【発明の実施の形態】本発明の請求項１に記載の発明は、辞書を用いて対象文書集合の各文書に出現する単語または単語の組の出現頻度や分布などの統計情報があらかじめ抽出されている文書集合に対して、各文書に付与された属性情報やユーザが入力した検索式などに基づいて文書の部分集合を特定する文書集合選定部と、各単語の対象文書全体における統計情報、および各文書ごとの当該文書に出現する単語とその統計情報を管理する単語統計情報管理部と、各単語の全文書および各文書ごとの統計情報を基に、特定された部分集合に出現する各単語の重要度を算出して重要度の順に整列する単語ランキング部とを備えたものであり、整列された単語群のうちの特定部分のみについて、単語もしくは単語とその重要度の組を抽出し、これを再利用可能な形で高速提示するという作用を有する。

【０００９】本発明の請求項２に記載の発明は、請求項１に記載の関連キーワード自動抽出装置において、特定された部分集合Ａに対して、これに含まれる部分集合Ｂが文書集合選定部により特定された場合に、部分集合Ａに含まれる文書群に出現する単語の統計情報と、部分集合Ｂに含まれる文書群に出現する単語の統計情報との差分を、部分集合Ｂにおける各単語の重要度に加味することで、部分集合Ｂに出現する各単語の重要度を算出して単語ランキングに反映するようにしたものである。

【００１０】本発明の請求項３に記載の発明は、請求項１または２に記載の関連キーワード自動抽出装置において、文書集合選定部に各文書の重みを付与する機能を設け、特定された文書集合の各文書に含まれる単語の重要度に当該文書の重みを加味することにより当該単語の重要度を算出して単語ランキングに反映するようにしたものである。

【００１１】本発明の請求項４に記載の発明は、請求項１乃至３のいずれかに記載の関連キーワード自動抽出装置において、対象文書集合全体において出現度合いが高頻度または低頻度である単語をあらかじめ定められた閾値を考慮して関連キーワード抽出の対象から除外することにより、再利用の際に有効性の高い単語のみが選別できるようにしたものである。

【００１２】本発明の請求項５に記載の発明は、請求項４に記載の関連キーワード自動抽出装置において、単語の長さなどその単語の特徴量に応じて除外のための閾値を変化させることにより再利用の際に有効性の高い単語のみが選別できるようにしたものである。

【００１３】本発明の請求項６に記載の発明は、請求項１乃至５のいずれかに記載の関連キーワード自動抽出装置において、単語の出現位置や出現する文脈の情報を管理する出現情報管理部を有し、単語の重要度にその単語の出現情報の種類に応じてあらかじめ定められた重みを加味することにより当該単語の重要度を算出して単語ランキングに反映するようにしたものである。

【００１４】本発明の請求項７に記載の発明は、請求項１乃至６のいずれかに記載の関連キーワード自動抽出装置において、単語の品詞など、各単語の属性情報を管理する言語属性管理部を有し、当該単語の属性に応じてあらかじめ定められた重みを加味することにより当該単語の重要度を算出して単語ランキングに反映するようにしたものである。

【００１５】本発明の請求項８に記載の発明は、請求項１乃至７のいずれかに記載の関連キーワード自動抽出装置において、抽出された単語同士、またはあらかじめ指定された単語群と抽出された単語との間の文字列としての包含関係を、定められた条件により判定する文字列包含関係判定部を有し、当該単語同士に文字列としての包含関係があると判定された場合に、指定された条件に従って、長単位の文字列のみ、もしくは短単位の文字列のみ、もしくは重要度の高い方の文字列のみ、もしくは短単位の文字列および長単位の文字列と短単位の文字列との差分の双方、のいずれかを選択することにより、再利用の際に有効性の高い単語のみが選別できるようにしたものである。

【００１６】本発明の請求項９に記載の発明は、請求項１乃至８のいずれかに記載の関連キーワード自動抽出装置において、単語の品詞など、各単語の属性情報を管理する言語属性管理部を有し、当該単語の属性や、指定された部分集合または文書全体における出現頻度、分布等を考慮することにより、抽出された単語を分類して提示できるようにしたものである。

【００１７】本発明の請求項１０に記載の発明は、請求項９に記載の関連キーワード自動抽出装置において、分類された単語群のそれぞれについて、その集合を代表する単語を付与する代表語付与部を設け、分類された単語群を代表する代表語群のみ、もしくは代表語と全ての単語を提示できるようにしたものである。

【００１８】本発明の請求項１１に記載の発明は、文書検索装置として、辞書を用いて対象文書集合の各文書に出現する単語または単語の組の出現頻度や分布などの統計情報があらかじめ抽出されている文書集合に対して、文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部と、文書検索部において検索された文書について、入力された検索式と文書との間の適合度を計算する文書ランキング部とを備えたものであり、文書ランキング部におけるランキング結果を関連キ

ーワード自動抽出装置へ送付し、また関連キーワード自動抽出装置からフィードバックされた関連キーワードを検索条件入力部へ入力するという作用を有する。

【００１９】本発明の請求項１２に記載の発明は、文書検索装置として、文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部とを備えたものであり、前記検索条件入力部は、ユーザが検索条件を入力するのに加えて関連キーワード自動抽出装置から送られてきた関連キーワードを検索条件として入力するという作用を有する。

【００２０】本発明の請求項１３に記載の発明は、文書検索システムとして、辞書を用いて対象文書集合の各文書に出現する単語または単語の組の出現頻度や分布などの統計情報があらかじめ抽出されている文書集合に対して、文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部と、文書検索部において検索された文書について、入力された検索式と文書との間の適合度を計算する文書ランキング部とを有して成る文書検索装置と、前記文書検索装置に接続された関連キーワード自動抽出装置とを備えたものであり、前記文書検索装置の文書ランキング部から出力されたランキング結果を関連キーワード自動抽出装置へ送付し、また関連キーワード自動抽出装置から文書検索装置の検索条件入力部へ関連キーワードをフィードバックしてキーワード検索を行なうという作用を有する。

【００２１】本発明の請求項１４に記載の発明は、請求項１３記載の文書検索システムにおいて、文書検索装置と関連キーワード自動抽出装置との間には文書集合選定部が設けられ、文書検索装置の文書ランキング部から出力されたランキング結果は文書集合選定部に送付されて文書の特定が行なわれ、前記関連キーワード自動抽出装置４８には、文書集合選定部４７が特定した文書の部分集合が入力されるようにしたものである。

【００２２】本発明の請求項１５に記載の発明は、請求項１３または１４記載の文書検索システムにおいて、関連キーワード自動抽出装置には、請求項１乃至１０のいずれかに記載の関連キーワード自動抽出装置が用いられるようにしたものである。

【００２３】本発明の請求項１６に記載の発明は、文書検索に必要な条件式を入力する検索条件入力部と、入力された検索条件にしたがって対象文書集合から文書の検索を行なう文書検索部とを有して成る文書検索装置と、前記文書検索装置に接続された関連キーワード自動抽出装置とを備えたものであり、前記文書検索装置の検索条件入力部は、ユーザが検索条件を入力するのに加えて関連キーワード自動抽出装置から送られてきた関連キーワードを検索条件として入力してキーワード検索を行なうという作用を有する。

【００２４】本発明の請求項１７に記載の発明は、請求項１６記載の文書検索システムにおいて、関連キーワード自動抽出装置には、請求項１乃至１０のいずれかに記載の関連キーワード自動抽出装置が用いられるようにしたものである。

【００２５】以下に、本発明の具体的な実施の形態について、添付の図面を参照して説明する。

【００２６】（実施の形態１）最初に、本発明の第１の実施の形態について説明する。図１は本発明の第１の実施の形態に係る関連キーワード自動抽出装置の構成を示したブロック図である。まず、対象となる文書集合１１に対し、辞書１２を利用して、前処理として動作する統計情報抽出部１３により、文書集合全体における単語の頻度・分布などの単語統計情報１４、および各文書ごとの当該文書に含まれる単語の統計情報である文書内単語統計情報１５を抽出しておく。図２（ａ）は単語統計情報の構造を示すテーブル構成図であり、図２（ｂ）は文書内単語統計情報の構造を示すテーブル構成図である。単語統計情報１４は、統計情報抽出部１３によって抽出された単語の統計情報を例えば図２（ａ）に示すようなテーブルとして格納する。このテーブルを利用することにより、例えば単語「インターネット」の全文書中での総出現頻度や出現文書数を高速に求めることができる。また、文書内単語統計情報１５は各文書ごとの単語の統計情報を例えば図２（ｂ）に示すようなテーブルとして格納する。これにより、例えば文書番号００１０には単語「インターネット」が５回、単語「WWW」が２回出現する、といった各文書ごとの統計情報を高速に求めることができる。

【００２７】関連キーワード自動抽出装置１６は、文書全体の単語統計情報１４および文書内単語統計情報１５を管理する単語統計情報管理部１７と、単語の重要度を算出する単語ランキング部１８と、対象文書の部分集合を特定する文書集合選定部１９と、文書集合選定部１９への選定条件を入力する手段である条件入力部２０とから構成される。

【００２８】かかる構成を有する関連キーワード自動抽出装置１６の動作について以下説明する。最初に、条件入力部２０に対して入力された条件により、文書集合選定部１９が文書集合を特定する。文書集合は、次の３種類の手段のいずれかまたはその組み合わせにより特定される。

（１）文書の属性により文書集合を特定する。この場合、文書集合選定部１９は文書の所属するジャンルなど、文書にあらかじめ付与された属性値により文書を選択する手段を有し、条件入力部２０により指定された属性値に合致する文書群を部分集合として採用する。

（２）検索式により文書集合を特定する。この場合、文書集合選定部１９は条件入力部２０で入力された検索式に適合する文書を特定する文書検索手段を有し、これを

利用して検索の結果得られる文書群を部分集合として採用する。なおその際、文書検索手段に検索式との適合度を判定して文書を適合度の順にランキングする機能があるならば、検索結果のうちの特定部分、例えば上位１０文書を部分集合として採用しても良い。

（３）ユーザにより指定された文書集合。この場合、文書集合選定部１９は条件入力部２０においてユーザが直接指定した（複数の）文書を部分集合として採用する。

【００２９】文書集合選定部１９は、以上により選定された文書集合を各文書を一意に決定する識別子の集合、例えば文書番号のリストとして単語統計情報管理部１７＊に渡す。単語統計情報管理部１７は、特定された文書集合に対して、文書ごとに文書番号から文書内単語統計情報１４を調べ、当該文書に出現する単語とそれぞれの文書内の出現頻度を得る。次に得られた単語すべてについて単語統計情報１５を調べ、当該単語の全文書における頻度や分布情報を得る。

【００３０】ここで得られた各種統計情報は単語ランキング部１８に渡され、各単語の重要度が算出される。ある単語Wの重要度Ｓ（W）は、例えば次のようにして算出することができる。

【数１】

$$S(W)=C*\sum_{j=0}^{n}\{TF_j(W)*IDF(W)\}*FN(W)$$

ただし

Ｃ　　：定数

ｎ　　：特定された文書集合に含まれる文書数

ＴＦｊ（Ｗ）：文書Ｄｊ における単語Wの出現頻度

ＦＮ（Ｗ）　：特定された文書集合中で単語Wを含む文書数

である。

【００３１】またＩＤＦ（Ｗ）は、単語Wのｉｄｆ値と呼ばれる指標であり、例えば以下の式により計算される。

ＩＤＦ（Ｗ）＝　１－ｌｏｇ（ＤＦ（Ｗ）／Ｎ）

ただし、

ＤＦ（Ｗ）：文書全体において単語Wが出現する文書数

Ｎ　　：全文書数

である。

【００３２】ＩＤＦ（Ｗ）は、単語Wがより多くの文書に出現する（すなわちより一般的な語である）場合にその値が小さくなる。これにより、対象文書全体において比較的よく出現する語の重要度を低く抑えることができる。さらにＦＮ（Ｗ）を考慮することで、特定された文書集合に多く現れる単語の重要度を高くでき、結果その特定文書集合に特徴的な語に高い重要度を与えることができる。なお、上記算出法において、ＴＦ（Ｗ）をその単語が含まれる文書の文書サイズ（文字数や含まれる単語の異なり数など）や単語の総出現頻度などで正規化してもよい。

【００３３】単語ランキング部１８は、特定された部分集合中の全文書に含まれる全単語について重要度計算を※行い、その後全単語を重要度の順に整列する。最後に、整列された単語群から特定部分、例えば上位１０単語を採用し、単語、もしくは単語とその重要度の組として提示する。なお、抽出の際に重要度だけでなく、重要度算出に利用した各種統計情報などを同時に提示してもよい。また、抽出された関連キーワードとその重要度の組を、例えばユーザの履歴として蓄積していくこともできる。このようにすることにより、ユーザの興味の範囲や嗜好などをキーワードとその重みのベクトルとして表現することが可能となり、このベクトルを他の操作、例えば文書集合の検索に利用するなど、広い応用が可能である。

【００３４】以上の計算式を利用すると、例えば図３に示す例のようにして関連キーワード自動抽出を行うことができる。この図３は関連キーワード自動抽出動作の処理手順の流れを示す図である。図３において、文書番号リスト３１が入力された単語統計情報管理部１７は、該当する文書番号（例えば００１０、０３４１等）に出現する単語およびその頻度を文書ごとに出力し、文書内単語統計情報３３，３４，３５を得る。同時に、ここで求められたすべての単語に対して、全文書中での統計情報３２を得る。次にこれらの統計情報３２、３３、３４、３５が単語ランキング部１８に渡される。単語ランキング部１８では、各種統計情報３２～３５を基に、例えば前記の式を利用して各単語の重要度を計算する。図３の場合だと、以下のようになる（ただし、Ｃを１、Ｎを１００００とする）。

ＩＤＦ（アプレット）＝　１－ｌｏｇ（８６／１００００）
＝　５．７５６
Ｓ（アプレット）　＝　２＊５．７５６＋６＊５．７５６）＊２
＝　９２．０９６
ＩＤＦ（インターネット）＝　１－ｌｏｇ（１１２９／１００００）
＝　３．１８１
Ｓ（インターネット）＝　（３＊３．１８１＋１＊３．１８１＋２＊３

．１８１）＊３
＝　５７．２５８
ＩＤＦ（ＣＧＩ）　＝　１－ｌｏｇ（７９／１００００）
＝　５．８４０
Ｓ（ＣＧＩ）　＝（４＊５．７５６）＊１
＝　２３．０２４
ＩＤＦ（ＷＷＷ）　＝　１－ｌｏｇ（６１５／１００００）
＝　３．７８９
Ｓ（ＷＷＷ）　＝（５＊３．７８９）＊１
＝１８．９４５
ＩＤＦ（ＪＡＶＡ）　＝　１－ｌｏｇ（１６１／１００００）
＝　５．１２９　　６
Ｓ（ＪＡＶＡ）　＝（６＊５．１２９＋３＊５．１２９＋３＊５．
１２９）＊３
＝　１８４．６４４
ＩＤＦ（ＳＵＮ）　＝　１－ｌｏｇ（３５／１００００）
＝　６．６５５
Ｓ（ＳＵＮ）　＝（６＊６．６５５）＊１
＝　３９．９３０
ＩＤＦ（スクリプト）＝　１－ｌｏｇ（８１３／１００００）
＝　３．５１０
Ｓ（スクリプト）　＝（５＊３．５１０）＊１
＝　１７．５５０

【００３５】単語ランキング部１８では以上のように求められた重要度により単語を整列し、整列後の単語リスト３７を得る。ここで、ランキングされた単語の上位３語を抽出するという指定になっているとすれば、単語リスト３７における上位３語である「ＪＡＶＡ」「アプレット」「インターネット」が関連キーワードとして抽出される。

【００３６】以上では辞書に登録された一単語を抽出の対象としてきたが、一般に単語だけでなく、単語の組でもよい。単語の組とは、名詞の連続により構成される複合語や、助詞「の」で結ばれる名詞の組、助詞「を」「が」で結ばれる名詞と動詞の組などを指す。これらの統計情報が単語と同様に事前に抽出できているのであれば、上記で示した手法がそのまま適用でき、単語の組を関連キーワードとして抽出することができる。

【００３７】なお、関連キーワード入力装置１６は、文書集合選定部１９および条件入力部２０を別構成としてもよい。特に文書集合選定部１９が検索式による文書検索手段を有する場合には、後出の図７に示すような別構成とすることで、文書検索装置による文書番号を入力として受け、出力される関連キーワードを文書検索装置の検索式入力部に反映させることができる。

【００３８】このように、本実施の形態によれば、対象となる文書のうちの一部である文書の部分集合が特定された際、当該部分集合に含まれる各文書に出現する各単語それぞれについて重要度を計算して重要度の順に整列し、整列された単語群のうちの一部を抽出して関連キーワードとすることで、動的かつ高速に対象となる文書の特性に即した関連キーワードを求めることができるという効果を持つ。

【００３９】また、上記のようにして得られた関連キーワードは、同一文書を対象とする文書検索装置への入力として利用することができ、その場合、対象文書の特性にあった的確なキーワードを再利用できるだけでなく、当該関連キーワードは必ず対象文書に含まれることが保証されるため、これを利用して検索した場合に必ず検索結果が得られるという効果も持つ。

【００４０】また、得られた関連キーワードを同一の対象文書集合または別の対象文書集合を対象とする文書検索装置への入力として利用することができ、その場合には、関連キーワード抽出の対象となった文書集合において特徴的であるキーワードをもとに、同一または別の文書集合を検索することができ、特に別の文書集合を検索対象とする文書検索装置の場合に、当該キーワードを異なった特性を持つ文書集合に対しても適用することができるという効果をもつ。

【００４１】また、抽出されたキーワードをユーザに提示して選択させるという構成とすることで、ユーザが再検索を実行する際、キーボードから再度検索条件を入力する代わりに、関連キーワードを、例えばマウスのクリックなど単純な操作で選択することが可能となり、再検索における操作を軽減して検索の効率を高めると同時に、検索の操作に不慣れなユーザでも簡単に利用できるという効果を持つ。

13

【００４２】また、抽出された関連キーワードにその重要度も付加して提示することにより、例えば検索条件との適合度を計算して文書をランキングするような文書検索装置において、検索条件中の各単語に対して重みを付与することができる文書検索装置であれば、抽出されたキーワードとその重要度をそのまま入力とすることで、より高精度の検索結果を得ることができるという効果を持つ。

【００４３】また、抽出された関連キーワードとその重要度の組を、例えばユーザの履歴として蓄積していくことにより、ユーザの興味の範囲や嗜好などをキーワードとその重みのベクトルとして表現することが可能となり、このベクトルを他の文書集合の検索に利用するなど、広い応用が可能であるという効果も持つ。

【００４４】（実施の形態２）次に、本発明の第２の実施の形態について実施の形態１に示したブロック図と同じ図１を利用して説明する。この第２の実施の形態では、文書集合選定部１９が２種類の文書集合Ａおよび文書集合Ｂを特定する。ここで、文書集合Ｂは文書集合Ａの部分集合となっている。例えば、ある検索式で検索を行った結果得られる文書集合Ａと、そのうちで関連する文書群としてユーザが指定した文書集合Ｂとが特定される場合や、文書の属性により特定された文書集合Ａと、その中でさらに検索式により絞り込まれた文書集合Ｂとが特定される場合などである。

【００４５】この場合、例えば以下の式により算出される単語の分布指標を当該単語の重要度に乗算するなどの手法により、単語の重要度を算出する。

ＤＩ（Ａ，Ｂ，Ｗ）＝　｛（ＮＡ／ＤＡ（Ｗ））＊（Ｄ＊Ｂ（Ｗ）／ＮＢ）｝

14

ただし、

ＤＡ（Ｗ）：部分集合Ａにおける単語Ｗの出現する文書数

ＤＢ（Ｗ）：部分集合Ｂにおける単語Ｗの出現する文書数

ＮＡ：部分集合Ａの総文書数

ＮＢ：部分集合Ｂの総文書数

【００４６】これは、部分集合Ｂにおいて高い頻度で出現し、かつ部分集合Ａにおける出現頻度が低いものほど高い値となる。上式において高い値となる語は部分集合Ａにおいて部分集合Ｂの弁別性に大きく寄与するものであり、部分集合Ｂをより特徴づけるキーワードであるといえる。例えば、図３に示す例において、文書番号リスト３１が部分集合Ｂであるとし、これを含む部分集合Ａ（総文書数１００とする）も同時に指定されている場合で、部分集合Ａ中の各単語の出現文書数が以下の通りであるとする。

ＤＡ（アプレット）　＝　１０
ＤＡ（インターネット）＝　２８
ＤＡ（ＣＧＩ）　＝　９
ＤＡ（ＷＷＷ）　＝　１４
ＤＡ（ＪＡＶＡ）　＝　２０
ＤＡ（ＳＵＮ）　＝　５
ＤＡ（スクリプト）　＝　１０

【００４７】この場合各単語の重要度Ｓ２（Ｗ）は、実施の形態１で説明した各単語の重要度Ｓ（Ｗ）に各単語の重みＤＩ（Ａ，Ｂ，Ｗ）を乗算した値となり、以下のように計算される。

Ｓ２（アプレット）　＝９２．０９６＊｛（１００／１０）＊（２／３）｝
＝　６１３．９７３

Ｓ２（インターネット）＝　５７．２５８　＊　｛（１００／２８）＊（３／３）｝
＝　２０４．４９３

Ｓ２（ＣＧＩ）　＝　２３．０２４　＊　｛（１００／９）＊（１／３）｝
＝　８５．２７４

Ｓ２（ＷＷＷ）　＝　１８．９４５　＊　｛（１００／１４）＊（１／３）｝
＝　４５．１０７

Ｓ２（ＪＡＶＡ）　＝１８４．６４４＊｛（１００／２０）＊（３／３）｝
＝　９２３．２２０

Ｓ２（ＳＵＮ）　＝　３９．９３０　＊　｛（１００／５）＊（１／３）｝
＝　２６６．２００

Ｓ２（スクリプト）　＝　１７．５５０　＊　｛（１００／１０）＊（１／３）｝
＝　５８．５００

となり、重要度の順に整列すると

Ｓ２（ＪＡＶＡ）　＝９２３．２２０
Ｓ２（アプレット）　＝　６１３．９７３
Ｓ２（ＳＵＮ）　＝　２６６．２００
Ｓ２（インターネット）＝　２０４．４９３
Ｓ２（ＣＧＩ）　＝　８５．２７４
Ｓ２（スクリプト）　＝　５８．５００
Ｓ２（ＷＷＷ）　＝　４５．１０７

の順となる。したがって、上位３語を関連キーワードとして抽出するのであれば、「ＪＡＶＡ」「アプレット」「ＳＵＮ」が関連キーワードとなる。

【００４８】上記の計算式は一例であり、部分集合Ｂにおいて高い頻度で出現し、かつ部分集合Ａにおける出現頻度が低いものほど高い値となるような他の計算式を利用してもよい。

【００４９】このように、本実施の形態によれば、特定された２種類の部分集合間における頻度分布の差異を考慮することにより、より高精度な関連キーワードを得ることができるという効果を持つ。

【００５０】（実施の形態３）次に、本発明の第３の実施の形態について実施の形態１に示したブロック図と同じ図１を利用して説明する。この第３の実施の形態では、文書集合選定部１９に各文書の重みを付与する機能を設ける。例えば、ユーザが文書を指定する場合に、各文書に対して関連度を指標として５段階の評価値を与える場合や、検索式による検索の結果得られる文書が検索式との適合度によりランキングされている場合に１位に１０点、２位に９点、といった重みを与える場合などである。単語ランキング部は各文書に付与された重みを、当該文書に含まれる単語に対して、例えば乗算するなどして加味し重要度算出を行う。なお、各文書に与える重みは負の値であってもよい。例えば、ユーザが文書を特定する際、関連する文書には２点、まったく関連しない文書には－１点を与える、という重み付与も許す。これにより、関連する文書にも関連しない文書にも含まれる（かつあまり一般的でない）語の重要度を低くすることができる。

【００５１】このように、本実施の形態によれば、特定した文書集合に含まれる各文書に対して重みを与えることにより、より重要な文書に含まれる単語ほど高い重要度となるような計算式とすることで、文書それぞれの重要度を勘案した高精度な関連キーワードが得られるという効果を持つ。

【００５２】（実施の形態４）次に、本発明の第４の実施の形態について説明する。図４は本発明の第４の実施の形態に係る関連キーワード自動抽出装置のブロック図である。この第４の実施の形態では、第１の実施の形態の構成に加えて閾値設定部２２を有して成り、この閾値設定部は単語統計情報管理部１７との間でデータの送受ができるようになっている。また、この実施の形態においては、単語統計情報管理部１７には閾値による単語除外機能が付与されている。かかる構成において、単語統計情報管理部１７は各単語の統計情報を出力する際、あらかじめ定められた閾値設定２２を参照し、極端に高頻度または低頻度の単語はその場で候補から除外して単語ランキング部１８に当該単語の情報を出力しない構成とすることができる。例えば、閾値１を「全文書の５０％以上に出現する単語」と設定し、閾値２を「１文書にしか出現しない単語」と設定することで、これらの単語が重要度算出に与える悪影響を事前に防ぐことができ、かつ処理の高速化を図ることができる。

【００５３】なおその際、単語の長さなど当該単語の特徴量に応じて、閾値を何種類かに設定してもよい。例えば、日本語の場合で「二文字以上の語は全体の５０％以上、一文字の語は全体の３０％以上」といった閾値設定を行うことで、各語の特性にあわせて除外する単語の範囲を設定する。

【００５４】このように、本実施の形態によれば、対象文書集合全体において出現度合いが高頻度または低頻度である単語をあらかじめ定められた閾値を考慮して除外することにより、キーワード抽出処理を高速化でき、かつ再利用の際に有効性の高い単語のみが選別できるという効果を持つ。

【００５５】（実施の形態５）次に、本発明の第５の実施の形態について説明する。図５は本発明の第５の実施の形態に係る関連キーワード自動抽出装置の構成を示すブロック図である。この第５の実施の形態に係る関連キーワード自動抽出装置は、第１の実施の形態において説明したような、文書全体の単語統計情報１４および文書内単語統計情報１５を管理する単語統計情報管理部１７、単語ランキング部１８、対象文書の部分集合を特定する文書集合選定部１９、および文書集合選定部１９への選定条件入力手段である条件入力部２０を有する基本構成に加えて、単語ランキング部１８と連動して単語の属性などの各種情報を利用することにより、抽出される関連キーワード群の質を向上させることを目的とするものである。図５において、符号２５は出現情報管理部、２６は単語属性情報管理部、２７は文字列包含関係判定部であり、これらの機能部は関連キーワード自動抽出装置２９に含まれて単語ランキング部と連動する。また２８は代表語付与部であり、この代表語付与部２８は単語ランキング部１８からデータを受けて関連キーワードを

出力する。また、関連キーワード自動抽出装置２９に対して、外部機能部として、対象文書集合１１からのデータを基に単語が出現する位置の情報を抽出する単語出現位置情報抽出部２３が設けられ、この単語出現位置情報抽出部２３からは出現位置情報２４が出力される。この出現情報は出現情報管理部２５へ送付される。

【００５６】かかる構成を有する本発明の第５の実施の形態について、その動作を説明する。この実施の形態の動作においては、まず対象となる文書集合１１に対し、辞書１２を利用して、前処理として動作する統計情報抽出部１３により、対象文書集合１１全体における単語の出現頻度・分布などの単語統計情報１４、および各文書ごとの当該文書に含まれる単語の統計情報である文書内単語統計情報１５を抽出しておく。同時に、必要があれば単語位置情報抽出部２３により、単語の出現位置情報２４も抽出しておく。図６は単語出現位置情報抽出部２３によって抽出された出現位置情報２４のデータ構造の一例を表すテーブル構成図である。出現位置情報は例えば図６に示すようなテーブルとして格納される。各文書ごとにその文書に出現する単語と出現位置（例えば文書の先頭からのバイトオフセット）、出現区分などが格納される。

【００５７】そして関連キーワード自動抽出動作に際しては、各単語に対して出現情報管理部２５に問い合わせを行い、当該単語の出現位置や出現文脈などの情報を得、これを重要度算出に加味する。例えば、検索対象とする文書すべてが、タイトル（または見出し）、サブタイトル、本文、といった要素から構成されている文書である場合、当該単語がこれら要素のうちいずれに含まれているかによって、

タイトルに含まれる場合には３点
サブタイトルに含まれる場合には２点
本文に含まれる場合には１点

といったような「重み」を各単語の重要度に乗算する、といった手法で重要度を算出する。

【００５８】あるいは、出現位置の情報を利用してもよい。例えば部分集合が検索式により特定される場合で、この検索式に含まれる単語が参照可能である場合、検索式に含まれる単語と、現在重要度計算の対象となっている単語との間の文字数が、

２文字以内なら３点
１０文字以内なら２点
１０文字以上なら１点

といったような「重み」を当該単語の重要度に乗算する、といった手法で重要度を算出することも可能である。

【００５９】また、本実施の形態の別の態様として、各単語に対して、単語属性情報管理部２６に問い合わせを行い、当該単語の品詞や分類など、その単語の属性を得、これを重要度算出に加味する。例えば、当該単語の品詞に着目し、

固有名詞ならば５点
普通名詞ならば４点
形容詞、形容動詞ならば２点
動詞、副詞ならば１点
その他自立語でないもの（助詞、助動詞など）ならば０点

といったような「重み」を各単語の重要度に乗算する、といった手法で重要度を算出することも可能である。

【００６０】また、本実施の形態の別の態様として、ある２つの単語間の文字列としての包含関係を判定する文字列包含関係判定部２７を用いて、抽出された単語同士、もしくはあらかじめ指定された単語群のうちの一単語と抽出された単語との間に包含関係があるか否かを判定し、包含関係があると判定された場合に、抽出する単語を制限する。ここであらかじめ指定された単語群とは、例えば部分集合の特定に検索式を利用した場合の検索式に含まれる単語などである。包含関係の判定においては、あらかじめ定められた設定により、以下の判定基準のいずれか一つ（または一つ以上）を満たす場合を包含関係と認定することができる。

（１）単語Ａと単語Ｂとが前方において一致しかつ単語Ａが単語Ｂより短い場合、（２）単語Ａと単語Ｂとが後方において一致しかつ単語Ａが単語Ｂより短い場合、（３）単語Ａが単語Ｂの部分でありかつ前方、後方ともに一致しない場合、（４）単語Ａと単語Ｂとの関係が（１）～（３）のいずれかを満たし、かつ単語Ｂの構成要素と完全に一致する場合、

【００６１】例えば、（１）の基準では「東京都」に対する「東京」が部分語と判定される。以下、同様にして、（２）の基準では「新発売」に対する「発売」が、（３）の基準では「大感謝祭」に対する「感謝」が、それぞれ部分語と判定される。（４）の基準は、英語における部分語判定の際に重要であり、この基準に従えば "artificial intelligence" に対して "art" や "tell" は部分語とはならないが、"artificail" や "intelligence" は部分語と判定される。

【００６２】上記基準により、部分語関係にあると判定された２つの語について、そのどちらを関連キーワードとして採用するかについても、以下のいずれかの基準（あらかじめ設定されているものとする）に従う。

（１）　長単位の単語を採用する
（２）　短単位の単語を採用する
（３）　重要度の高い単語を採用する
（４）　短単位の単語および長単位の単語と短単位の単語との差分を採用する

【００６３】例えば、単語「東京都」が重要度１０で、単語「東京」が重要度７でそれぞれ抽出され、かつ両者に部分語関係が成立した場合、（１）の基準に従うと文字列として長い「東京都」が採用され、（２）の基準に

従うと文字列として短い「東京」が採用され、（３）の基準に従うとより重要度の高い「東京都」が採用されることになる。（４）の基準は、例えば単語 "artificial intelligence" と "artificial" との間に部分語関係が成立した場合に、"artificial" および "intelligence" を関連キーワードとして採用するものであり、主に英語文書において効果的である。

【００６４】あらかじめ指定された単語群との間に部分語関係が成立する単語の場合、（３）以外の手法が利用できる。その場合、「短単位（もしくは長単位）であれば関連キーワードとして採用しない」といった処理となる。抽出された単語同士に部分語関係が成立する場合には、いずれの手法も利用可能である。

【００６５】また、本実施の形態の別の態様として、抽出された関連キーワード群を、各語の属性や統計情報を利用して分類して提示する。語の属性として品詞を利用すると、例えば固有名詞とそれ以外に分類して提示することができる。あるいは、語の属性としてシソーラス辞書を利用し、各語をシソーラスにおける分類に対応する形で分類して提示することも可能である。また、統計情報を利用した分類とは、例えば特定された文書集合における各語の出現文書数により分類する手法などがあげられる。その場合、例えば「出現文書数が文書集合の８割以上であるか否か」といった基準で分類することで、その語が再検索に利用される際の絞り込みの効果を事前に確認することができる。なお、分類にあたり語の属性としてシソーラス辞書を利用する場合、分類された単語群に対して、シソーラスの上位ノードに相当する語を代表語として与え、単語群をその語で代表させることも可能である。同様に、単語の統計情報１４を利用する場合には、分類された単語群において、例えば最も出現頻度の高い語を代表語として採用してもよい。

【００６６】このように、本実施の形態によれば、単語が出現した位置の情報を利用することで、文書の構造や単語間の距離の情報を考慮した関連キーワードの抽出が行なえ、高精度な関連キーワード抽出が可能となるという効果を持つ。

【００６７】また、単語の品詞など、各単語の属性情報を考慮することにより、各属性の特徴に応じた関連キーワードの抽出が行なえ、高精度な関連キーワード抽出が可能となるという効果を持つ。

【００６８】また、単語間の文字列としての包含関係を考慮することにより、同じような意味や用途である単語を排除して関連キーワードの抽出が行なえ、関連キーワード全体としての冗長性を抑えることができるという効果を持つ。

【００６９】また、抽出された関連キーワードを分類し、必要があれば各分類に対応する代表語を設定することで、抽出されたキーワードの一覧性や傾向、再利用の際の有効性などをあらかじめ確認して関連キーワードの抽出が行なえ、関連キーワードとしての使いやすさを向上することができるという効果を持つ。

【００７０】（実施の形態６）次に、本発明の第６の実施の形態について説明する。図７は本発明の第６の実施の形態に係る文書検索装置の構成およびこれと関連キーワード自動抽出装置とを組み合わせて実現した文書検索システムの構成を示すブロック図である。この文書検索装置４１は、前記第１，第２，第３、第４または第５の実施の形態に係る関連キーワード自動抽出装置と連携して動作するものである。

【００７１】本実施形態における文書検索装置４１は、文書検索に必要な条件式を入力する検索条件入力部４４と、入力された検索条件にしたがって文書の検索を行なう文書検索部４５と、文書検索部４５において検索された文書について入力された検索式と文書との間の適合度を計算する文書ランキング部４６とを有して成る。この文書検索装置４１は、連携して動作する関連キーワード自動抽出装置４８と同一の対象文書集合１１を検索対象とするものであり、単語統計情報抽出に利用するのと同一の辞書１２を利用して、あらかじめ索引生成部４２により作成された文書検索用の索引４３を利用して検索を行う。また、本実施形態における関連キーワード自動抽出装置４８は、文書集合選定部４７を別構成としたものであり、関連キーワード自動抽出装置４８には、文書集合選定部４７が特定した文書の部分集合の各要素に対応する文書の識別子の集合（一意である文書番号のリストなど）が入力される。

【００７２】以上の構成を備えた本実施の形態について、その動作を説明する。最初に検索条件入力部４４に入力された検索条件をもとに、文書検索部４５が検索用索引４３を参照して検索条件に適合する文書を特定する。ここで得られた文書集合をそのまま検索結果文書５０としてもよいが、さらに文書ランキング部４６において、入力された検索式と文書との間の適合度を計算して適合度の高い順に文書を整列したものを検索結果とする、といった構成にしてもよい。こうして得られた検索結果の文書集合５０は、ユーザに検索結果として返すのと同時に、文書集合選定部４７に渡される。文書選定部４７では、文書ランキング部４６から渡された文書集合のすべてまたは一部を関連キーワード自動抽出装置４８への入力として採用する。文書が適合度の順にランキングされているのであれば、検索結果の文書集合のうち例えば上位１０文書を選定する、という構成にしてもよい。また、あらかじめ文書ごとに付与された属性情報が利用できるのであれば、これを利用して特定の属性値を持つ文書のみを選定する、という構成としてもよい。

【００７３】文書集合選定部４７により特定された文書の部分集合は関連キーワード自動抽出装置４８に送られ、前記第１，第２，第３，第４または第５の実施の形態に示したような手順で関連キーワード群４９を抽出す

る。こうして得られた関連キーワード群４９は検索条件入力部４４に戻され、ユーザに提示される。ユーザは提示された関連キーワード群から必要なものを選択して新たな検索条件とし、再度検索を実行させることができる。これにより、本実施の形態によれば、関連キーワード自動抽出装置によって上記のようにして得られた関連キーワードは、同一文書を対象とする文書検索装置への入力として利用することができ、その場合、対象文書の特性にあった的確なキーワードを再利用できるだけでなく、当該関連キーワードは必ず対象文書に含まれることが保証されるため、これを利用して検索した場合に必ず検索結果が得られるという効果も持つ。

【００７４】（実施の形態７）次に、本発明の第７の実施の形態について説明する。図８は本発明の第７の実施の形態に係る文書検索装置の構成およびこれと関連キーワード自動抽出装置とを組み合わせて実現した文書検索システムの構成を示すブロック図である。この文書検索装置５１は、第６の実施の形態に係る文書検索装置４１と同様、前記第１，第２，第３、第４または第５の実施の形態に係る関連キーワード自動抽出装置と連携して動作するものである。

【００７５】本実施形態における文書検索装置５１は、文書検索に必要な条件式を入力する検索条件入力部５４と、入力された検索条件にしたがって文書の検索を行なう文書検索部５５とを有して成る。本実施形態における文書検索装置５１は、連携して動作する関連キーワード自動抽出装置５２とは異なる対象文書集合５６を検索対象とするものであり、文書検索部５５が対象文書集合５６に接続される構成となっている。なお、その検索手法についての詳細は問わない。

【００７６】以上の構成を備えた本実施形態における動作について、以下説明する。最初に指定された条件にしたがって関連キーワード自動抽出装置５２が動作し、関連キーワード群５３を出力する。文書検索装置５１における検索条件入力部５４は、関連キーワード群５３を入力としてユーザに提示し、ユーザは提示された関連キーワードのうち必要なもののみを選択して、検索対象となる対象文書集合５６に対する検索を実行し、検索結果文書５７を得ることができる。

【００７７】このように、本実施の形態によれば、関連キーワード自動抽出装置５２によって得られた関連キーワードを同一の対象文書集合または別の対象文書集合を対象とする文書検索装置５１への入力として利用することができ、その場合には、関連キーワード抽出の対象となった文書集合において特徴的であるキーワードをもとに、同一または別の文書集合を検索することができ、特に別の文書集合を検索対象とする文書検索装置の場合に、当該キーワードを異なった特性を持つ文書集合に対しても適用することができるという効果をもつ。

【００７８】【発明の効果】以上説明したように、本発明によれば、関連キーワード自動抽出装置を、文書の部分集合を特定する文書集合選定部と、対象文書全体または個々の文書ごとに出現する単語とその統計情報を管理する単語統計情報管理部と、文書の部分集合に出現する各単語の重要度を算出して重要度の順に整列する単語ランキング部ととにより構成したため、文書全体、および特定された文書部分集合における各単語の統計情報を高速に求めることが可能であり、特定された文書集合に出現する各単語を、その重要度に基づいて高速にランキングし、その一部を関連キーワードとして提示することができる。

【００７９】また、前記構成に加えて、単語の属性情報や文書内の出現位置を管理する手段などを設けることにより、単語の重みを変化させ、あるいはランキング後の単語群から特定の条件を満たす単語を削除することで、抽出される単語群の関連語としての精度を向上させることができる。また、抽出された単語群を、語の属性や統計的性質により分類することで、よりわかりやすい関連キーワード提示を行なうことができる。

【００８０】さらに、関連キーワード自動抽出装置と連携した文書検索装置を含む文書検索システムを構成し、抽出された関連キーワードを入力として再利用することにより、抽出された関連キーワードが対象文書の特性に合ったものであり、かつ検索対象が同一の文書群であるならば、そのキーワードによって検索結果が少なくとも１件以上得られることが保障されるため、効率的かつ容易に再検索を行なうことができる等の効果が得られる。

【図面の簡単な説明】

【図１】本発明の第１乃至第３の実施の形態に係る関連キーワード自動抽出装置の構成を示すブロック図

【図２】（ａ）　前記実施の形態における単語統計情報の構造を示すテーブル構成図

（ｂ）　前記実施の形態における文書内単語統計情報の構造を示すテーブル構成図

【図３】前記実施の形態における関連キーワード自動抽出動作の処理手順の流れを示す図

【図４】本発明の第４の実施の形態に係る関連キーワード自動抽出装置の構成を示すブロック図

【図５】本発明の第５の実施の形態に係る関連キーワード自動抽出装置の構成を示すブロック図

【図６】前記実施の形態における単語出現位置情報抽出部によって抽出された出現位置情報のデータ構造の一例を表すテーブル構成図

【図７】本発明の第６の実施の形態に係る文書検索装置の構成構成およびこれと関連キーワード自動抽出装置とを組み合わせて実現した文書検索システムの構成を示すブロック図

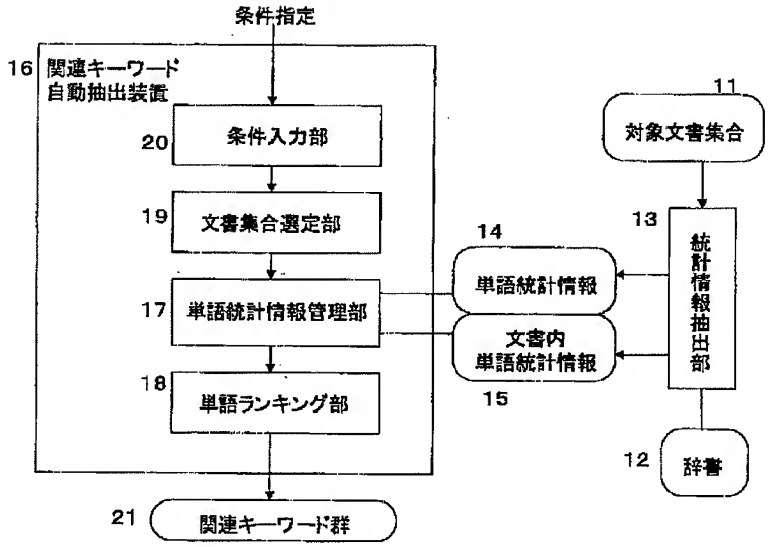
【図８】本発明の第７の実施の形態に係る文書検索装置の構成およびこれと関連キーワード自動抽出装置とを組み合わせて実現した文書検索システムの構成を示すブロ

ック図

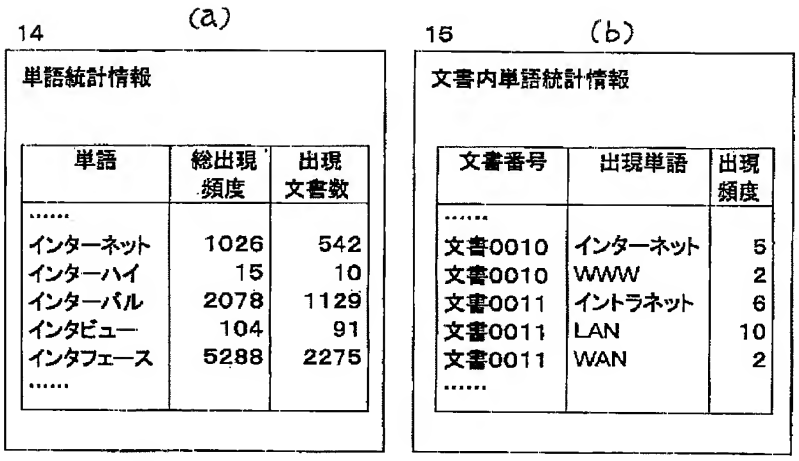
【符号の説明】

１１、５６　対象文書集合
１２　辞書
１３　統計情報抽出部
１４　単語統計情報
１５　文書内単語統計情報
１６、２９、４８、５２　関連キーワード自動抽出装置
１７　単語統計情報管理部
１８　単語ランキング部
１９　文書集合選定部
２０　条件入力部
２１、４９、５３　関連キーワード群
２２　閾値設定
２３　単語出現位置情報抽出部
＊２４　出現位置情報
２５　出現情報管理部
２６　単語属性情報管理部
２７　文字列包含関係判定部
２８　代表語付与部
４１、５１　文書検索装置
４２　索引生成部
４３　検索用索引
４４、５４　検索条件入力部
４５、５５　文書検索部
４６　文書ランキング部
４７　文書集合選定部
５０、５７　検索結果文書

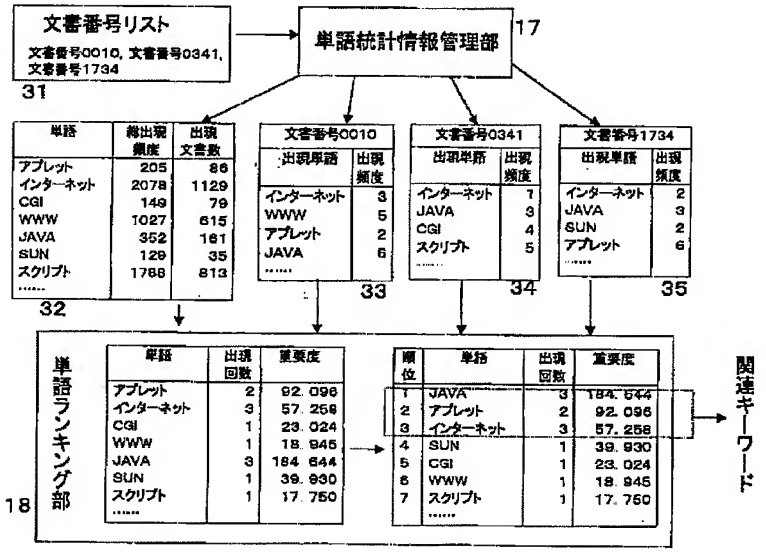
【図１】



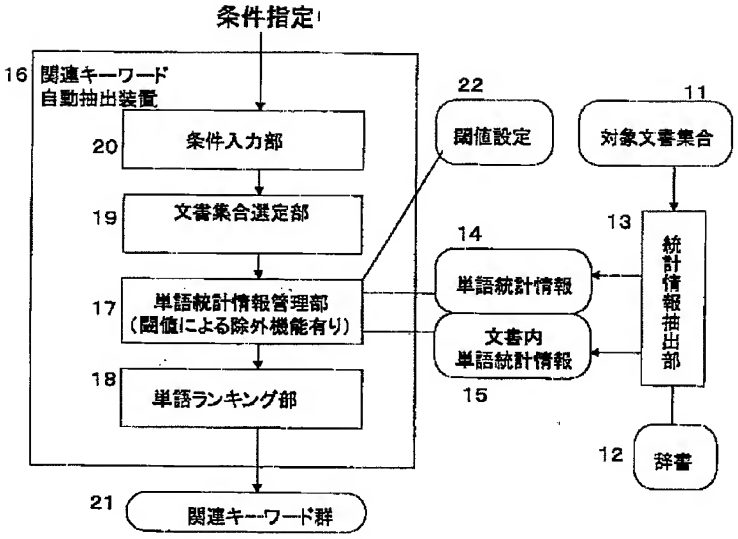
【図２】


14
(a)
単語統計情報
単語
総出現頻度
出現文書数
インターネット 1026 542
インターハイ 15 10
インターバル 2078 1129
インタビュー 104 91
インタフェース 5288 2275
15
(b)
文書内単語統計情報
文書番号
出現単語
出現頻度
文書0010 インターネット 5
文書0010 WWW 2
文書0011 イントラネット 6
文書0011 LAN 10
文書0011 WAN 2

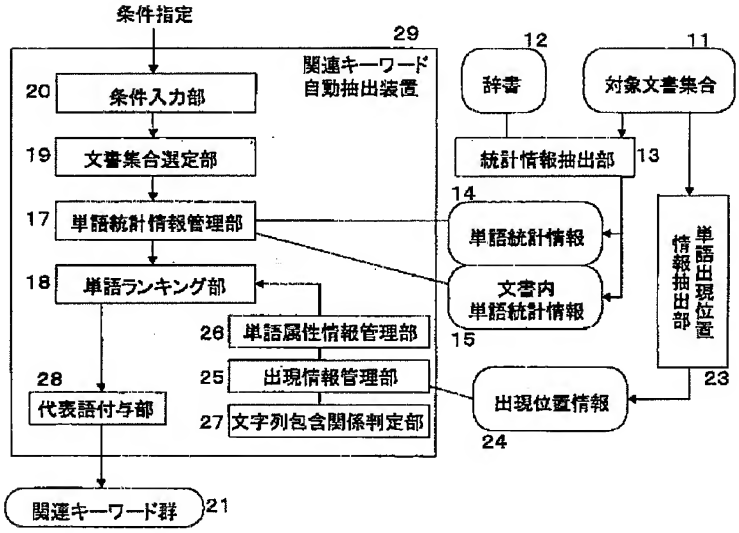
【図３】


文書番号リスト
文書番号0010, 文書番号0341, 文書番号1734
31
単語統計情報管理部
17
32
33
34
35
単語ランキング部
18
関連キーワード

【図４】


条件指定
16 関連キーワード自動抽出装置
20 条件入力部
19 文書集合選定部
17 単語統計情報管理部（閾値による除外機能有り）
18 単語ランキング部
21 関連キーワード群
22 閾値設定
11 対象文書集合
13 統計情報抽出部
14 単語統計情報
文書内単語統計情報
15
12 辞書

【図５】


条件指定
29 関連キーワード自動抽出装置
20 条件入力部
19 文書集合選定部
17 単語統計情報管理部
18 単語ランキング部
26 単語属性情報管理部
25 出現情報管理部
27 文字列包含関係判定部
28 代表語付与部
関連キーワード群 21
12 辞書
11 対象文書集合
統計情報抽出部 13
14 単語統計情報
文書内単語統計情報
15
単語出現位置情報抽出部 23
出現位置情報 24

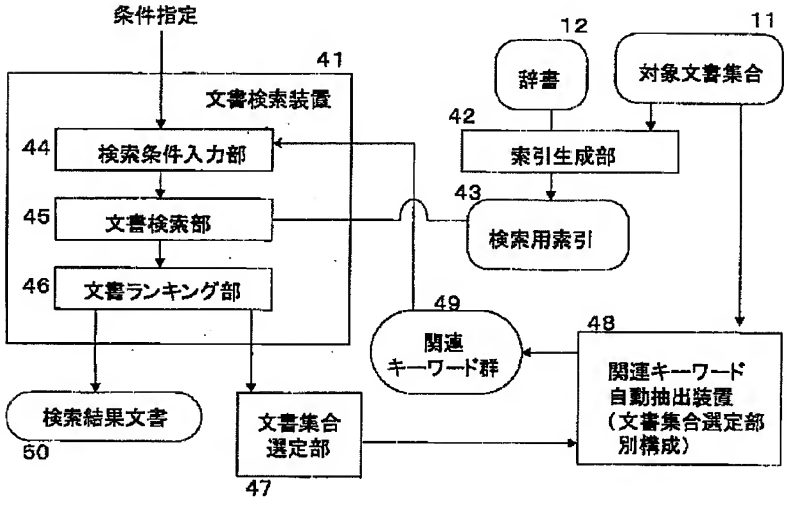
【図６】

24

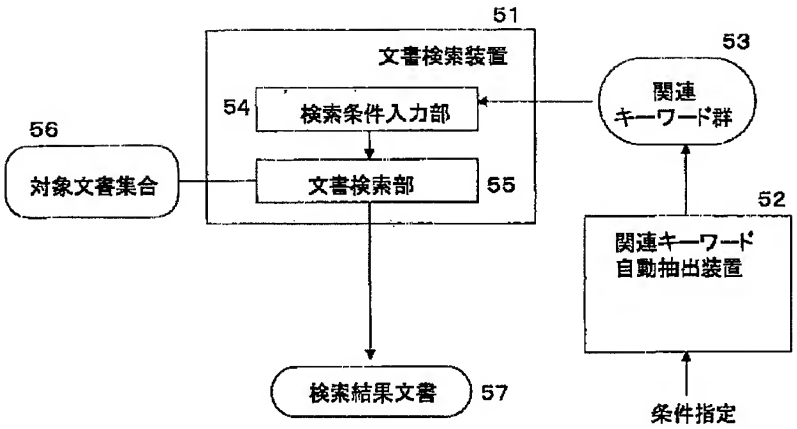
単語出現位置情報

文書番号	出現単語	出現位置	出現区分
……			
文書0010	インターネット	10368	タイトル
文書0010	WWW	10384	タイトル
文書0010	近年	10390	本文
文書0010	インターネット	10396	本文
文書0010	成長	10412	本文
……			

【図７】



【図８】



フロントページの続き

(72)発明者　野　本　昌　子
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　稲　葉　光　昭
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　福　重　貴　雄
大阪府門真市大字門真1006番地　松下電器産業株式会社内